Computing controversy: Formal model and algorithms for detecting controversy on Wikipedia and in search queries

AUTHOR(S):

Zielinski, Kazimierz; Nielek, Radoslaw; Wierzbicki, Adam; Jatowt, Adam

CITATION:

Zielinski, Kazimierz ...[et al]. Computing controversy: Formal model and algorithms for detecting controversy on Wikipedia and in search queries. Information Processing & Management 2018, 54(1): 14-36

ISSUE DATE:

2018-01

URL:

http://hdl.handle.net/2433/230645

Information Processing and Management 54 (2018) 14–36

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Computing controversy: Formal model and algorithms for detecting controversy on Wikipedia and in search queries

Kazimierz Zielinski*, Radoslaw Nielek, Adam Wierzbicki, Adam Jatowt

Polish-Japanese Academy of Information Technology, Warsaw, Poland, Kyoto University, Kyoto, Japan

ARTICLE INFO

Article history:
Received 22 January 2017
Revised 11 August 2017
Accepted 23 August 2017
Available online 22 September 2017

Keywords:
Controversy
Wikipedia
Web search

ABSTRACT

Controversy is a complex concept that has been attracting attention of scholars from diverse fields. In the era of Internet and social media, detecting controversy and controversial concepts by the means of automatic methods is especially important. Web searchers could be alerted when the contents they consume are controversial or when they attempt to acquire information on disputed topics. Presenting users with the indications and explanations of the controversy should offer them chance to see the "wider picture" rather than letting them obtain one-sided views. In this work we first introduce a formal model of controversy as the basis of computational approaches to detecting controversial concepts. Then we propose a classification based method for automatic detection of controversial articles and categories in Wikipedia. Next, we demonstrate how to use the obtained results for the estimation of the controversy level of search queries. The proposed method can be incorporated into search engines as a component responsible for detection of queries related to controversial topics. The method is independent of the search engine's retrieval and search results recommendation algorithms, and is therefore unaffected by a possible filter bubble.

Our approach can be also applied in Wikipedia or other knowledge bases for supporting the detection of controversy and content maintenance. Finally, we believe that our results could be useful for social science researchers for understanding the complex nature of controversy and in fostering their studies.

1. Introduction

Controversy abounds in the world. At times it seems that almost any subject can be a source of controversy since it is virtually impossible for everyone to agree on any subject of interest. However, such assumption is a fallacy: by focusing on the presence of minor disagreements, we ignore the existence of an organized, encyclopedic collection of agreed-upon facts. Still, it is quite clear that controversial topics and controversial information exist, and compared with non-controversial topics or information, they are usually more prominent and easily noticeable in media and on the Web. Even topics that are well supported by facts and evidence may not constitute a guaranteed consensus. People tend to argue about opinions, interpretations, and points of view. These disagreements are not always counterproductive; they may allow participants to

* Corresponding author.
E-mail address: kazimierz.zielinski@pjwstk.edu.pl (K. Zielinski).

http://dx.doi.org/10.1016/j.ipm.2017.08.005

adjust their points of view or to inspire a search for truth or common ground. Yet, controversy can also impede or prevent cooperation. This problem has been recognized and addressed by communities that generate peer-produced content, such as Wikipedia, which attempts to limit the impact and spread of controversy and improve the article editing process. Therefore, detection of controversy may be crucial for ensuring high-quality peer-produced content and more balanced viewpoints.

While browsing the Web in search for information about a topic of interest, individuals may find content that supports only one point of view. The threat of receiving biased information or misinformation is even more serious with the pervasive personalization of search engines and the associated filter bubble effect (Nguyen, Hui, Harper, Terveen, & Konstan, 2014; Pariser, 2011). Unless users persevere in their search for acquiring more information to corroborate and to find information from diverse viewpoints, they are likely to learn about only one-side of any argument regarding the topic of their interest. Consequently, a person may make wrong decisions as a result of being exposed to misleading (or one-sided) information about a controversial topic, without even realizing that controversy exists. If the person was made aware of the controversy, she/he would be able to make more informed and better decisions based on more comprehensive knowledge about the topic. The truism 'forewarned is forearmed' also applies to the context of Web searches.

With the above in mind, a timely warning about the existence of controversy could be beneficial for users who may be unaware of it before they trust incomplete or biased information on important topics. In our study, we focus on the two most widely used "starting points" for users who are looking for information on the Internet: Wikipedia and search engines. With close to 500 million unique visitors per month Wikipedia is immensely popular knowledge base that has powerful educational impact on our society. Often users use it as a springboard when researching variety of topics including complex ones in scientific domains or more trivial ones like celebrity information. Wikipedia is also commonly used for supporting various data processing tasks.

Search engines are gateways to the Internet. They assist users in locating any desired information or websites. As such, it would be beneficial to searchers if they got alerted on topic controversy when interacting with search engines – before even seeing the returned results. In this work we consider only search queries and not the results returned by search engines to avoid the need to account for the search engine type and limitations associated with personalizing the search progress.

1.1. Research objectives

Given the described importance and complexity of the concept of controversy, we identified the following research objectives.

While prior work has already demonstrated methods of identifying controversy, there is no unified framework that would define the problem and represent it formally. We believe that a conceptual model of controversy is necessary. It would provide basis for computational approaches towards building applications aiming at detecting, estimating and understanding controversy in texts. Such a model would be useful for establishing detection methods, but it could also allow us to compare existing approaches.

Given the importance of Wikipedia and search engines in acquiring information and learning, one would expect a strong support for online users to prevent acquisition of one-sided information. However, no such support has been explicitly offered so far, besides solutions that rely on manual detection of controversy or focus on related tasks (e.g., content relevance estimation). To implement such functionalities, the development of effective algorithmic approaches for treatment of controversy is required.

1.2. Contributions

In this work, we make the following contributions:

Definition and formal representation of controversy: We propose a definition of controversy to serve as the basis for our research. This definition is based on the social sciences and epistemology. It may also be useful for researchers in computer science, as it draws a line between controversial and non-controversial content and topics. The main advantage of this definition is that the level of controversy can be quantified. Moreover, our formal model classifies errors as having three main sources; such classification is indispensable when attempting to predict controversy.

Estimating controversy of Wikipedia articles and categories: We develop new methods for identifying controversial articles on Wikipedia by analyzing the sentiments of discussions on the article talk pages. Using these detection methods, we can pinpoint the sections of articles that generate the greatest controversy. Based on the proposed classifiers of Wikipedia article controversy, we can estimate the levels of controversy associated with topical categories on the English-language Wikipedia. This approach can be then used to predict the level of controversy in new articles within a category.

Method for estimating controversy levels of search queries for offering early warning about high controversy: Our third contribution is a method for estimating controversy and generating early warnings about high controversy associated with the topics of Web search queries. This method is based on a ranking of the level of controversy characterizing Wikipedia content categories. We then consider the similarity of a query to Wikipedia content categories associated with known controversy levels. This prediction tool can be used by searching Web users to receive early warnings prior to the submission of a search query to a Web search engine. Most importantly, the proposed method only depends on the user's query and information from Wikipedia, not on the search engine or its search results. Because our method is independent of the search engine's algorithm, it also works independently of the filter bubble. Note that two different users may receive

different, personalized results to the same query. This situation is most significant for queries related to controversial topics, because a user may receive search results that reflect one side of the controversy (and another user may only learn about the other side). In contrast, our method produces results that are independent of the personalized recommendation algorithm that selects a user's search results.

The remainder of this article is organized as follows: in the next section, we review the abundant literature on the topic of controversy and focus on work relevant to defining controversy and on algorithmic methods for controversy detection or prediction. In Section 3, we propose our definition of controversy. In Section 4, we describe the Wikipedia dataset used in our research. In Section 5, we introduce and evaluate our methods for predicting controversy of Wikipedia articles. In Section 6, we evaluate and rank the level of controversy of topical categories in Wikipedia. Section 7 introduces our method for predicting the controversy in Web queries. Finally, in Section 8, we draw conclusions and discuss the future work.

2. Related work

2.1. Controversy in social science

Controversy is a ubiquitous phenomenon. Since the times of ancient Greece to the 20th century, the study on the nature of controversy was a domain of great philosophers. It is usually defined as public argument/discussion about a topic that has supporters on one side and many people who strongly disagree with them or are somehow shocked by this issue. With respect to time, controversy is characterized by the fact that no agreement or consensus is reached on the subject for a prolonged period of time. Controversies range from fierce polemics to polite and orderly discussions. What makes controversy research an exciting topic is that controversies usually stem from issues of high social relevance.

Today, the study of controversy is carried out mainly by philosophers, sociologists, and information scientists. Controversies are known to exist in all areas where there are disputes relevant to current societies, including economy, politics, art, religion, science, gender, sexual orientation, race, ethnicity, culture, immigration, education and many other areas. Controversy, by its very nature, is related to conflict. As early as in 1956, the functions of social conflicts were described by Coser (1956), who considered the role of conflict in establishing and maintaining group identities and relationships between them. Conflicts that arise from the frustration of specific demands are more persistent and lead to aggressive behavior based on strong sentiments. In those individuals who are deeply involved in the controversy, feelings of attraction as well as hostility are likely to arise. Sometimes parties view themselves as representatives of collectives or groups who are fighting not for themselves but rather for the goals and ideals of the group. The elimination of personal motivations tends to intensify conflict. Conflict also leads to the formation of coalitions and associations between previously unrelated parties.

Some of Coser's observations regarding social conflicts are particularly meaningful in the Internet controversies of today, wherein we find that ideological conflicts, for example, are often particularly difficult and lead to ongoing controversies. This is because in such cases the cause is not related to any real-world fact or incident, but rather to differing world views or identities, which correspond to Coser's social groups. Currently, we also frequently observe the expression of strong sentiments and the formation of coalitions between parties in a controversy.

Koutra, Bennett, and Horvitz (2015) examined the browsing behavior of searchers on the controversial and polarizing topic of gun control, and in this study focused on the influence of a single disruptive and shocking news event. The authors found that, overall, most people use the Web to access information with which they agree. When an external event threatens to directly influence users, only then do they explore content outside their filter bubble. Otherwise, most users appear "narrow-minded" with respect to the number of visited Web domains, and these domains present mostly one side of the issue, which is typical in ideological conflict (and associated controversy), as defined by Coser.

In 1976, Gurr and Duvall (1976) formulated a formal theory of political conflict. This theory, developed in the pre-Internet era, focuses on the physical confrontation between social collective actors and the associated material destruction incurred. While the authors' main purpose in developing the theory was to describe social phenomena such as rebellions and other physical confrontations between organized collectives, it is applicable to Internet conflicts and controversies in the sense that the magnitude of a conflict still depends on individual potentials for action, collective dispositions toward action, and organizational strength – the latter to a lesser extent because in the Web2.0 era we are dealing with informal groups and coalitions of Internet users.

2.2. Role of controversy in media

Controversy plays very important role in media as it typically increases media popularity and coverage. Media programs like reality TV were developed for the purpose of sparking controversies by creating moral panic, as described in Biltereyst (2004). The MIT Center for Civic Media developed a Controversy Mapper, wherein major news stories are reverse-engineered to visualize the spread of ideas and the change in media frames over time, and to identify whose voices are dominating a discussion. In their investigations of a major national controversy in the media, the MIT researchers found that broadcast media continues to be important as an amplifier and gatekeeper of news, but that it is susceptible to media activists working through participatory media to co-create news and influence the framing of major controversies. Mejova, Zhang, Diakopoulos, and Castillo (2014) found that, in general, when it comes to controversial issues, the use of negative affect and biased language is prevalent, while the use of strong emotion is somewhat tempered. The authors also

found many differences between news sources, thus enabling the identification of controversy based on a comparison of how the subject is portrayed by different media groups. An unsupervised method for identifying controversial semantic categories in online media was proposed by De Clercq et al. (2014). The authors used DBpedia to aggregate concepts into categories and then identified categories characterized by significant deviations in sentiment across different online media.

2.3. Controversy in social media

As we will detail later, controversy needs appropriate community. Research on mining antagonistic communities in social networks was undertaken by Lo, Surian, Prasetyo, Zhang, and Ee-Peng (2013) who modeled direct antagonistic sub-communities using existing positive and negative links between members of *Eopinions* and *Facebook*. In Twitter, a widely used participatory media, controversies receive wide coverage very quickly. However, Smith, Zhu, Lerman, and Kozareva (2013) found that Twitter is primarily used for spreading information to like-minded people rather than for debating issues. Users are more likely to rebroadcast information than to respond to a communication by another user. Individuals typically take a position on an issue prior to making posts about it and are unlikely to change their opinion. Yardi and Boyd (2010) observed group dynamics on Twitter between pro- and anti abortion rights groups in relation to a certain event and discovered both homophily and heterogeneity in conversations about abortion. Although people today are exposed to broader viewpoints than they were previously, they are limited in their abilities to engage in meaningful discussion.

In their research on controversies involving known entities in Twitter, Pennacchiotti and Popescu (2010) used timely and historical scores to classify controversy. The authors found that most controversies on Twitter relate to micro-events (e.g., TV shows, award shows, or sport events). Garimella, Morales, Gionis, and Mathioudakis (2015) used the social media network structure on Twitter to develop a network-based controversy measure and found content features to be relatively unhelpful in this task. Later they focused their research on exploring the topics of discussion on Twitter and understanding which ones are controversial (Garimella, Mathioudakis, Morales, & Gionis, 2016).

2.4. Controversy in Wikipedia

Wikipedia provides a knowledge base that has been exploited by researchers from various scientific domains. The research trends in areas such as information retrieval, natural language processing, and ontology building had been reviewed and described recently by Mehdi, Okoli, Mesgari, Nielsen, and Lanamaki (2017).

The first paper addressing the problem of cooperation and conflict between editors on Wikipedia was published in 2004 by Viégas, Wattenberg, and Dave (2004). The authors used the history of edits to develop a tool for visualizing patterns of conflict. Based on similar meta-information (i.e., number of edits, unique editors, and anonymous edits), Stvilia, Twidale, Smith, and Gasser (2005) measured article quality. Buriol, Castillo, Donato, Leonardi, and Millozzi (2006) studied conflicts in Wikipedia by focusing on entries characterized by a series of edits reverting them to previous versions between authors.

Wikipedia edits are organized around the concept of *revision*. An editor will make changes to an article and publish a new version (aka revision). A current revision may always be reverted to an earlier version (by anyone) and the entire history of revisions is accessible on Wikipedia. Repetitive mutual reverts have led to edit wars and typically indicate the existence of controversy (or at least an argument about some facts). An in-depth study of edit wars has been conducted by Sumi, Yasseri, Rung, Kornai, and Kertész (2011a) and Sumi, Yasseri, Rung, Kornai, and Kertesz (2011b). Not all edit wars are necessarily non-constructive. In their study, Yasseri, Sumi, Rung, Kornai, and Kertész (2012) identified mutual reverts that have led to consensus as well as those that have remained in a state of permanent controversy. Later Yasseri, Spoerri, Graham, and Kertész (2014) used edit wars to present, visualize and analyze topical overlaps of controversies in 10 different language versions of Wikipedia. The authors concluded that there are both global controversial topics with cross-culture resonance and topics with more narrow interest limited to language communities or geographical areas.

To detect controversial articles, researchers can access a broad spectrum of meta information regarding editorial behaviors. Vuong et al. proposed two controversy-ranking models for Wikipedia articles, drawn from the history of collaboration and edits (Vuong et al., 2008). Kittur, Suh, Pendleton, and Chi (2007) proposed a very good method based on edit dynamics and talk pages. Recently, Rad and Barbosa (2012) compared different methods for detecting controversy (some of which are based on the properties of collaboration networks), which constituted the basis for one of our recent research efforts *Predicting Controversy of Wikipedia Articles Using the Article Feedback Tool* (Jankowski-Lorek, Nielek, Wierzbicki, & Zieliński, 2014). Focusing on the collaboration history of pairs of editors Sepehri-Rad and Barbosa introduced a model to predict controversy in a supervised way (Sepehri-Rad & Barbosa, 2015).

Sentiment analysis has proven to be very effective in many applications, including predicting the outcome of political elections (Sang and Bos, 2012; Tumasjan, Sprenger, Sandner, and Welpe, 2011; Wawer & Nielek, 2008), stock companies valuations (Feldman, Rosenfeld, Bar-Haim, and Fresko, 2011; Wawer & Nielek, 2009), and review analyses (Nielek, Wawer, and Wierzbicki, 2010; Wang, Zhu, & Li, 2013). However, it has not typically been used in Wikipedia research. Among more than 2000 articles focusing on Wikipedia, Okoli, Mehdi, Mesgari, Nielsen, and Lanamaki (2012) identified only a few that had applied sentiment analysis. Wikiganda (Chandy, 2008) investigated the occurrence of propaganda in controversial Wikipedia articles by measuring sentiment in articles with a single revision. Ferschke, Gurevych, and Chebotar (2012) analyzed the sentiment of the Talk Pages on Simple English Wikipedia with respect to dialog acts.The goal was to identify successful patterns

of collaboration that increased the article quality. Laniado, Kaltenbrunner, Castillo, and Morell (2012) analyzed the emotional styles of Wikipedia editors. Finally, Dori-Hacohen and Allan (2013) used sentiment as basis for detecting controversy on the Web and compared these results with a method based on a tagged corpus of articles from Wikipedia. Wikipedia has received much research attention, particularly in the fields of social behavior and content quality. In contrast to studies of conflict, Borzymek, Sydow, and Wierzbicki (2009), Turek, Wierzbicki, Nielek, Hupa, and Datta (2010), and Wierzbicki, Turek, and Nielek (2010) investigated collaboration and teamwork on Wikipedia social network.

Wikipedia has been successfully used as data source for retrieving semantic information to improve results from search engines and to categorize texts. Milne and Witten (2008) measured semantic relatedness by the hyperlink structures of Wikipedia articles. They counted *term frequency-inverse document frequency* (tf-idf) of links weighted by the probability of each link and then computed the relatedness of each article. Hajian and White (2011) created a multi-tree for each entity in the Wikipedia categories network, combined them, and used a multi-tree similarity algorithm to compute the similarity of the entities. Recently, Han (2013) proposed a method for measuring semantic similarity that uses Wikipedia as an ontology.

In addition to semantic similarity, the Wikipedia category graph (WCG) has been used in research to improve ad hoc document retrieval (Kaptein, Koolen, & Kamps, 2009), identify document categories (Schonhofen, 2006), and acquire knowledge (Nastase & Strube, 2008). Medelyan, Milne, Legg, and Witten (2009) published an extensive overview of research which mined Wikipedia. In 2006, Voss (2006) coined a term *collaborative thesaurus* for the Wikipedia category structure. The structured form of Wikipedia categories allowed for the automated learning of ontology (Yu, Thom, & Tam, 2007). Kittur, Chi, and Suh (2009) used a WCG of annotated data to detect contentious topics in Wikipedia. Recently, Biuk-Aghai, Pang, and Si (2014) attempted to visualize human collaboration in Wikipedia by visualizing WCG subtrees as simple trees.

In their search for controversial topics in Wikipedia articles, Borra et al. (2015) used language agnostic programming to develop a tool they called Contropedia. It assigns controversy scores to every Wiki link. These scores are then presented graphically as a controversy dashboard showing both the Wiki link score and its associated timeline of edits.

2.5. Credibility and controversy

With the advent of Web2.0, online evaluation had become an important feature in many applications involving information. In order to assess the credibility of information on the Web, a number of evaluation systems were developed and much research had been devoted to the design of credibility assessment systems. Kąkol, Jankowski-Lorek, Abramczuk, Wierzbicki, and Catasta (2013) studied the influence of subjectivity and bias in credibility ratings and found their presence to be a key design issue in Web credibility systems. The concept of controversy has been studied with respect to evaluation systems as an important influencing factor for improving overall credibility. The existence of subjectivity and mutual dependency between reviewer bias and object controversy is broadly recognised. Lauw et al. developed a reinforcement-based model to quantify bias and controversy within an evaluation system (Lauw, Lim, & Wang, 2006; 2008).

2.6. Controversy in Web search

Ennals, Trushkowsky, and Agosta (2010b) developed a Dispute Finder browser extension that finds and highlights text snippets that correspond with known disputes. A database of known disputes is manually maintained and a textual entailment algorithm is used to find snippets that correspond with a known dispute. Initially, the authors populated the database from websites like Politifact and Snopes, and subsequently they developed a method for identifying disputes on the Internet based on some English text patterns found to be typical in these disputes (Ennals, Byler, Agosta, & Rosario, 2010a). Kawahara, Inui, and Kurohashi (2010) focused on the recognition and bird's eye view presentation of contradictory and contrastive relationships between statements related to a topic (search query) and those expressed on Web pages (search engine results). Their method is based on predicate-argument language structures in Japanese, but is believed to be adaptable to other languages. Finding contradictions through the sentiments expressed by viewers was attempted by Tsytsarau, Palpanas, and Denecke (2010). Their method required the topic in question to have a number of online reviews or opinions, from which a sentiment could first be aggregated and then used to identify a controversy. Dori-Hacohen and Allan classified controversies on Internet pages, based on their similarity to Wikipedia articles known to be controversial. Initially, the authors based their classifications on a manually annotated set of articles (Dori-Hacohen & Allan, 2013), but then they generalized their method for automatic classification (Dori-Hacohen & Allan, 2015). The authors' recent research (Dori-Hacohen, Jensen, & Allan, 2016) proposed a new stacked model and a corresponding method of classification of controversial pages by creating subnetworks of topically related pages. They achieved improvement in AUC of the developed classifier when compared with the classifier based on a subnetwork of random pages, thus proving that controversy exists in topical neighborhoods. Jang and Allan (2016) addressed the unavailability of edit history for some pages by smoothing from the scores of the neighbors with more established edit history, which helped them improve the binary controversy classification. Jang, Foley, Dori-Hacohen, and Allan (2016) investigated the probability of controversy in a document and developed a language model for controversy. Yamamoto (2012) developed a Web query support system by collecting disputed sentences related to search queries and then providing the user with some of the most typical and relevant disputed sentences to enhance users' awareness of suspicious statements. Despite these advances, more problems arise with respect to controversial queries and controversial results if our goal is to provide assistance to search engine users. As described by Dori-Hacohen, Yom-Tov, and Allan (2015), challenges arise regarding the scope and context of queries, scientific correctness

vs. popular beliefs, the need for moral judgment when faced with several possible correct answers, and the cultural and social settings of the person seeking information. Going further beyond controversy detection Dori-Hacohen proposed research on automatic stance detection (Dori-Hacohen, 2015).

3. Formal model of controversy

3.1. Formal model of controversy

Many researchers have studied controversy but, surprisingly, no formal computational model for detecting content controversy has yet been proposed or widely accepted. The main reason for this is the issue's complexity caused by multiple aspects of the problems. Diverse methods, heuristics and algorithms for detecting controversy proposed in recent years applied to different types of objects (starting from sentences and ending with products or video clips). However, it is difficult to create a comprehensive theoretical framework for them. On the other hand such a model is crucial for the development and evaluation of algorithms for detecting controversy. In this section we outline the model and discuss potential difficulties in using it to shape and predict controversies.

The formal model of controversy proposed in this section cannot be calculated *per se* without adding content- and community-dependent methods of calculating (or estimating) empirical distribution of community opinions about evaluating subject and selecting the appropriate metric for measuring disagreement (it might be as simple as Leik's ordinal consensus or as complex as application of clustering and earth mover distance proposed by Rafalak, Deja, Wierzbicki, Nielek, & Kakol, 2016). It is also worth to notice that proposed model is universal and can be applied to any type of objects and their features – e.g. controversy about the trustworthiness of web pages or the taste of ice cream.

The Merriam-Webster Dictionary defines controversy as an *"argument that involves many people who strongly disagree about something: strong disagreement about something among a large group of people"*. From this definition, we can extract the two building blocks of controversy: a large group of people and the opinions or evaluations they express. If we combine these two aspects with the fact that controversy must be about something (e.g., text, event, or picture) we obtain a triad composed of the object, a group of people, and their opinion about the given object. Separately, neither a group of people (community) nor an object is sufficient to generate controversy. Formally, controversy can be defined as a function of three variables:

$$f(ob, com, E^{ob}_{com}) \rightarrow \{uncontroversial, controversial\}, \tag{1}$$

where

O – *set of objects*
$ob \in O$ – *single object of controversy (e.g., web page, picture, opinion, etc.),*
C – *set of communities,*
$com \in C$ – *community,*
E^{ob}_{com} – *empirical distribution of opinions given by members of community* ***com****for object* ***ob****.*

Object *ob* is controversial for a given community *com* only if there is strong disagreement in its evaluation E^{ob}_{com}. Inter-rater agreement measures (e.g., Fleiss' kappa or intra-class correlation coefficient) or dispersion measures of results (e.g., standard deviation) are typically used to assess agreement but are not suitable for use with a Likert-type ordinal scale (Jamieson et al., 2004). It is typically used to evaluate objects on the Web. Historically, the first measure tailored to deal with ordinal scales was the Leiks ordinal consensus, introduced in 1966 by Leik (1966). Other metrics were proposed by Tastle and Wierman (2007) and Van der Eijk (2001). Yet another approach is to measure the distance between the distribution of opinion E^{ob}_{com} and the reference distribution to identify a controversial class (e.g., u-shaped distribution). The earth mover's distance and the Bhattacharyya distance are two of many possible metrics for discrete distributions. Although controversy may vary in intensity, for simplicity we assume that a proposed function returns only one of two values – either a given object *ob* is controversial or not – but the same approach may be used for functions incorporating more shades of gray.

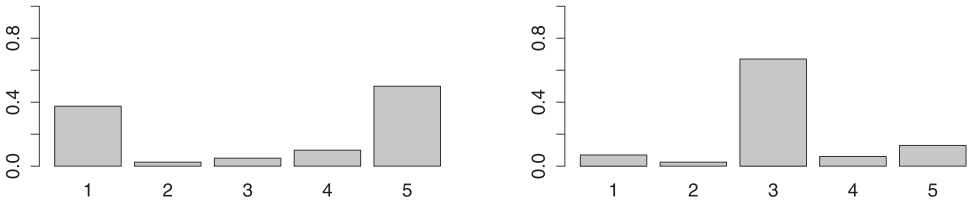
Eq. (1)can be simplified by removing objects *ob* and *com* from the function, because in our calculation we only use the object *ob* evaluations and not the object or community themselves. As a result we obtain the following:

$$f(E^{ob}_{com}) \rightarrow \{uncontroversial, controversial\}. \tag{2}$$

Following the above definitions, it becomes clear that the controversy cannot be analyzed without regard to the community. One consequence of this is that various communities may have different controversies on the same object. In Wikipedia, an obvious example of these disjunctive communities can be seen in its different language versions. An example of how the controversy on one topic changes across Wikipedia language versions is described in the Section 5. Fig. 1 depicts two actual examples of evaluation distributions for different articles in the English Wikipedia. From these graphs, we see that Justin Bieber seems to be, at least for the Wikipedia community, much more controversial than Alexandre Dumas.

3.2. Meta approaches for estimating object controversy

Evaluating the controversy on a given object is relatively easy as long as all the data is available, namely, a list of all the members of the community and their honest evaluation of object *ob*. Unfortunately, this is rarely the case. Usually



(a) Justin Bieber (b) Alexandre Dumas

Fig. 1. Distributions of votes on trustworthy rating obtained via Article Feedback Tool for two articles.

some data is missing and must be compensated for in some way. Below we present three typical real-world scenarios and a brief analysis of the potential sources of errors. We note that there is one more important aspect of computer-based controversy. Both in the model we present in Section 3 and in our discussion below, we assume that neither the object nor the community changes over time. Web pages evolve, people change their opinions, and communities grow. Each of these processes may strongly influence the calculation results of Eq. (1) and impose yet other limitations on controversy prediction methods.

3.2.1. Estimation by similarity to other objects

In this scenario, we have high-quality data regarding the controversy about object ob_a for a community com_i but no information about object ob_b's evaluation, for which we want to estimate the controversy. So, we can try to measure the similarity between objects ob_a and ob_b and, if we find them to be very similar, we can assume that the results of the controversy function defined in Eq. (1) will be the same for both. The main problem here is that calculating the similarity between real objects can be done in many different ways and we do not know a priori which of these methods is most accurate in approximating the controversy function values.

Our attempts to apply this approach to texts by using tf-idf for measuring the similarity of texts, however, except in some rare cases, were not successful. This is not surprising if you consider that the presence of even a single word in a lengthy text might spark controversy. For other types of objects, such as pictures or video, this issue becomes even more challenging.

3.2.2. Estimation of distribution and community

A common approach to studying controversy in the computer sciences is to ask people to label whether a given object is controversial and then to use these labels for further study (e.g., applying machine-learning algorithms and building models (Dori-Hacohen & Allan, 2013; Dori-Hacohen et al., 2015)). Researchers, being aware that controversy is an inherently fuzzy concept, collect many evaluations for each object and then calculate the inter-rater reliability. When people are asked about the controversy of a given object, they have to judge the object and estimate two additional variables required to calculate Eq. (1): *com* and E^{ob}_{com}. People tend to overestimate the popularity of their own opinion in a community and following *availability heuristic* think of themselves as a typical community member. Both of these potential error sources may be somewhat controlled by gathering answers from many people instead of from a single individual, but as yet it is unknown whether this is more effective than collecting subjective evaluations about an object.

3.2.3. Bias sources

There are quite a few services on the Internet on which people can rate objects, including movies, books, restaurants and others. Even the English Wikipedia had a mechanism for rating articles (see Section 4.1). In that case we know the community and the object, and we have evaluations of that object. Therefore in theory we should be able to precisely calculate the controversy function (Eq. (1)). In reality it is not that simple.

Of the millions of Wikipedia readers, only a tiny fraction evaluate a given article. Moreover, this small sample might not be representative of the whole community since it is self-selected. People tend to evaluate objects that inspire their intense emotions (either positive or negative). Therefore, the opinions of the silent majority are usually not well represented in collected evaluations. Rafalak et al. (2016) found the distributions of evaluations to have relatively robust sampling bias with respect to socio-demographic variables. The over-representation of a particular point of view in our sample might also be due to intentional manipulation – even a small minority of well-organized users can strongly influence the distribution of evaluations.

4. Datasets

In this section we describe Wikipedia derived datasets used for our research.

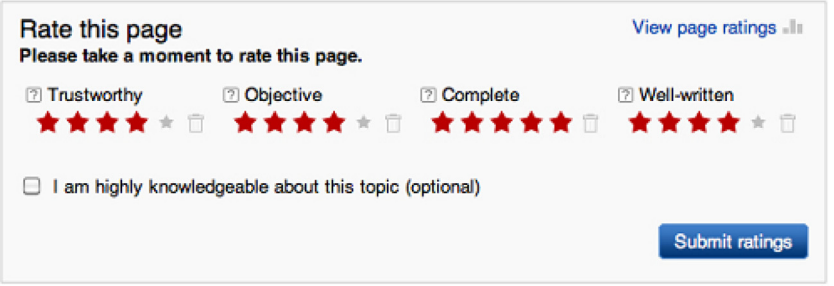


Fig. 2. Article Feedback Tool v4.

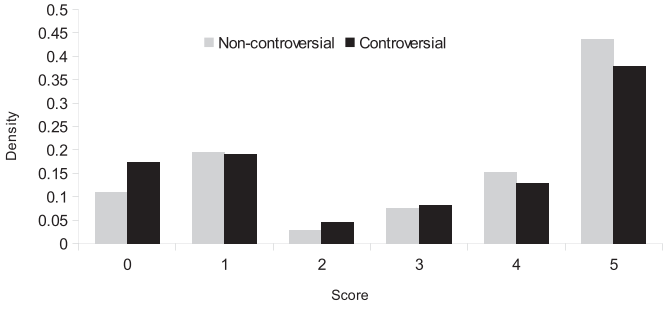


Fig. 3. Histogram of rating scores (0–5) of trustworthy dimension per class.

4.1. Article Feedback Tool

Article Feedback Tool (AFT) – a Wikimedia survey for article feedback – engaged readers in the assessment of article quality. AFT was used on Wikipedia in five versions and was recently discontinued. In our research, we use AFT v4, which allowed readers of the English Wikipedia to rate articles on four aspects, regarding whether they were perceived as being trustworthy, objective, complete, and well-written. Users were presented with a survey (Fig. 2) at the end of every article and could submit their ratings for each dimension on a star scale of 0 to 5. In this research, we considered only the *trustworthy* dimension of AFTv4 in order to create universal controversy detection classifiers that can be applied without directly asking users about the objectivity of the text. Our original AFTv4 data dump contained over 11 M article ratings based on 5.6 M different revisions of more than 1.5 M distinct articles collected between July 2011 and July 2012.

4.2. Articles dataset

The primary source of Wikipedia articles tagged as controversial is a list composed by Wikipedia administrators. For the English Wikipedia, this "list of controversial issues" (the official list title) contains 963 articles, less than 0.04% of all articles. This very small number of controversial articles may indicate that only a fraction of articles is correctly tagged (an example of one decidedly controversial article that is not on the list is given in Section 5). A careful examination of the list also reveals that many of these articles are no longer controversial (although they were controversial in the past and may be so again in the future). We tested many alternative ways for collecting a list of controversial articles (including a dataset shared by Dori-Hacohen & Allan, 2013), but decided to continue to use a list composed by the Wikipedians, as it was the easiest to replicate for other researchers and, at the same time, represented the closest example of the wider community understanding of controversy.

From a total number of 963 controversial articles, we chose 219 that had at least three evaluations in the AFT dataset. Since the application of machine learning techniques requires both positive and negative examples (preferably in similar quantities), we also randomly selected non-controversial articles in a way that maintained the same distribution of length (in characters) in both classes. The final dataset thus contained records of 438 articles, both controversial and non-controversial. Fig. 3 presents ratings of the trustworthy dimension from the AFTv4 database for each of the classes. We can observe that controversial articles had more 0-star ratings and fewer 5-star ratings. As displayed in Fig. 4, controversial articles had generated many more ratings than non-controversial ones. Almost 70% of the non-controversial articles had no more than 20 ratings, whereas 56% of the controversial articles had more than 100 ratings. Controversial articles seem to attract more voters.

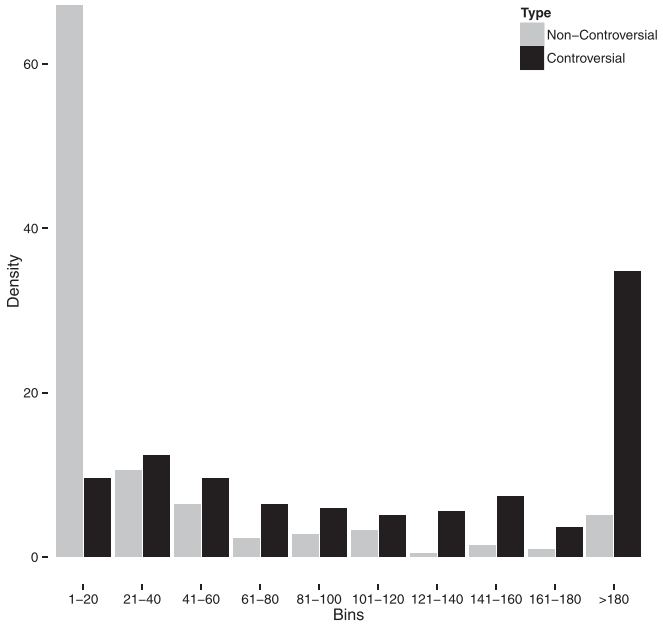


Fig. 4. Histogram of number of ratings per article.

4.3. Sections dataset

To determine whether it is possible to automatically identify controversial issues within an article, we randomly selected 512 sections from the talk pages of controversial articles and manually annotated them, using binary values (1 for controversial and 0 for non-controversial). Surprisingly, only 19.5% of the sections turned out to be controversial.

5. Detecting controversy of Wikipedia articles

To manage the editing of controversial articles, Wikipedia has established a set of rules. Although these rules can vary among different language versions, in principle they are based on limiting the number of users who are authorized to edit. On one hand, restraining the editing of controversial articles reduces edit wars, but on the other hand it slows the publication of quality content. Early detection of controversy may help to limit the cost and impact of edit wars and increase the reliability of Wikipedia articles. On Wikipedia, a topic that is controversial in one language version may not be controversial in another. One example of this is an article about the historic bombing event of the Polish town of Wielun by the German Luftwaffe at the outbreak of WWII. The controversy between some Polish and German historians can only be found in the English Wikipedia. A much longer article covering this event exists on the Polish Wikipedia, which is not characterized by any controversy, while there is no such article at all on the German Wikipedia. This example fits our formal model of controversy and reflects the dependency of controversy on the community.

Below, we present three classifiers for the detection of controversy on Wikipedia. We are proposing the first two (Sections 5.1 and 5.2), whereas we are using the other one (Section 5.3.1), which was introduced by Kittur et al. (2007), as a baseline for comparison.

5.1. AFT-based controversy detection

For every article in our dataset we computed the frequency of each of five available AFT ratings, the Leik's dispersion measure (Leik, 1966), and the total number of ratings. We used the results as features in a machine learning algorithm (Random Forest). We found the most important features to be the number of 1-star ratings, the number of 2-star ratings, and the total number of ratings.

5.2. Emotion polarity-based controversy detection

Editors use more sentiment-loaded language on talk pages when referring to Wikipedia articles. Most talk pages are organized into sections and comments, whereby the sections correspond to particular discussion topics. We calculated the frequencies of negative/neutral/positive words for each talk page/section/comment and used these as input for classifiers. Then, we used the random forest algorithm to train the emotion polarity classifier (EP), using both the annotations and sentiment measures.

Table 1
Performance comparison in AFT dataset.

	AFT	EP	Kittur	Combined
F-measure	80.37%	69.05%	81.48%	**84.10%**
Precision Controversial	79.96%	67.66%	79.77%	**83.14%**
Precision Non-controversial	80.78%	70.78%	83.42%	**85.11%**
ROC-AUC	0.88	0.73	0.89	**0.91**

5.3. Baseline

As comparative baseline for our algorithms, we used the method proposed by Kittur et al. (2007). In addition, we determined whether combining or not these method with our algorithms increased its performance.

5.3.1. Meta information

Kittur et al. proposed the meta classifier (Kittur et al., 2007), which we used as one baseline for our study. It had been compared with five other classifiers in a previous study (Rad & Barbosa, 2012). We chose this method because of its good performance and the easiness of dataset construction. Based on a statistical analysis of the revision histories for articles and their talk pages, we found seven features out of 30 to be the most important:

- number of revisions of the discussion page
- number of minor edits of the discussion page
- number of unique editors of the discussion page
- number of revisions of the article
- number of unique editors of the article
- number of revisions of the discussion page by anonymous editors
- number of revisions of the article by anonymous editors

Based on the above, we trained a classifier to distinguish between controversial and non-controversial articles.

5.4. Performance of classifiers

The baseline classifier we used to evaluate the performance of our proposed classifiers is described in Section 5.3.1 above. We used the random forest (Dori-Hacohen & Allan, 2015) implementation in R for machine learning. Due to the stochasticity of this algorithm, we ran the learning process for each model 50 times with different seed values and then we calculated the average of the confusion matrix.

5.5. Detecting controversial articles

Table 1summarizes the performance of the algorithms we describe in this section. The algorithm proposed by Kittur et al. (2007) yields the best results but its overall advantage over our AFT algorithm is negligible (slightly more than 1% for the F-measure (accuracy) and 0.01 for the receiver operating characteristic-area under the curve (ROC-AUC)). The EP method performs substantially worse (F-measure of 69.05%). The Combined method (Kittur, AFT and sentiment) works best for our dataset and reached 85.11% precision for non-controversial articles. Thus we can claim that the AFT-based method, which follows our formal model of controversy performs approximately as good as Kittur's baseline method.

For many applications, the misclassification cost is not symmetric, i.e., identifying a truly non-controversial article as controversial is less costly than doing the opposite. Fig. 5 shows the ROC for all the tested algorithms. The AUC peaks at 0.91 for the algorithm that combines the methods described in this paper, but they are also quite high for the Kittur and AFT algorithms (0.88 and 0.89, respectively). Again, the difference between the Kittur and AFT algorithms is very slight, but we can observe an interesting difference in the shapes of the ROC curves for these algorithms. AFT works better for lower false positive ratios and Kittur surpasses AFT at higher false positive ratios. This may explain why combining these two methods improves the performance. In general, the steep rise in the ROC functions for all the algorithms (except emotion polarity) indicates that by accepting slightly more false positives, for which misclassification is not costly, almost all controversial articles can be detected.

5.6. Classification performance based on emotion polarity

Fig. 6shows the classification performance of the EP model based on the sections on the discussion page. In Fig. 5 we saw that the EP-based classifier showed a significantly weaker performance than the other two classifiers, when applied to entire talk pages, with its AUC being equal to 0.73. We increased the granularity level of the sentiment calculations and applied them to the discussion sections rather than to the talk pages. As there are no meta data or AFT ratings available at the sections level, the only performance comparison we can make is with the performance in the AFT dataset. When

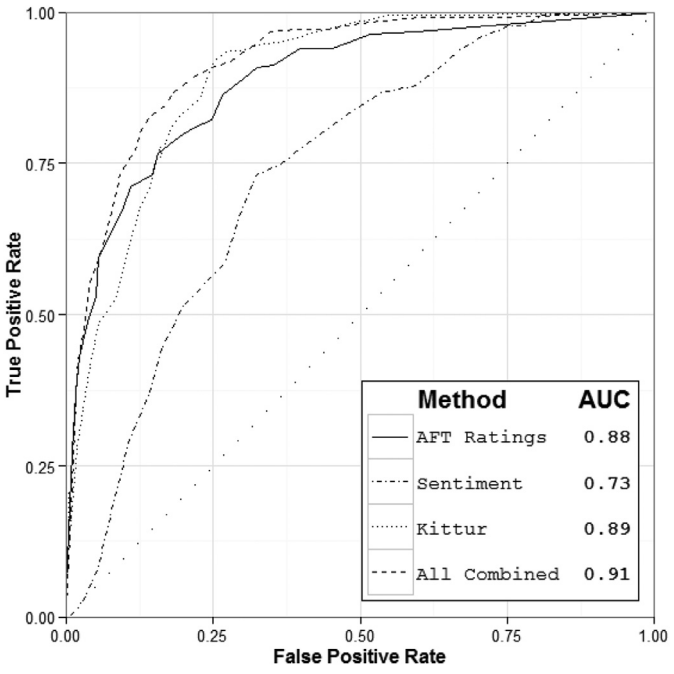


Fig. 5. ROC of classifiers in article dataset.

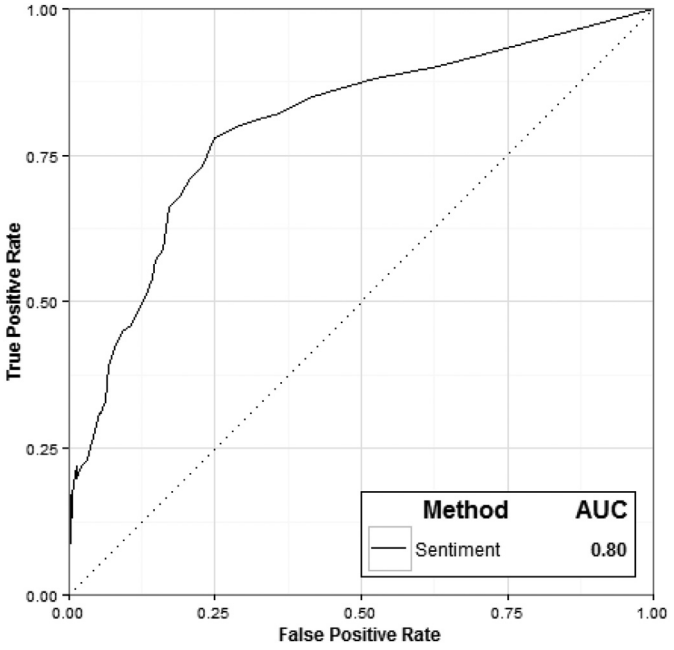


Fig. 6. ROC of sections classifier.

we look at the performance of the EP-based classification of sections, we see that its AUC is equal 0.80, which is better than when it is applied to entire talk pages (AUC of 0.73). The reason for the performance improvement is because even for articles of known controversy, the controversy is identified in only 19.5% of the discussion sections. The remaining 80.5% of the sections are not characterized by controversy. We can say that the controversy is “concentrated” in a few discussion threads, and becomes “diluted” by non-controversial threads when we calculate joint sentiment scores for entire talk pages that include both controversial and non-controversial sections. If we then train the classifier on such diluted measures, the classification performance inevitably decreases.

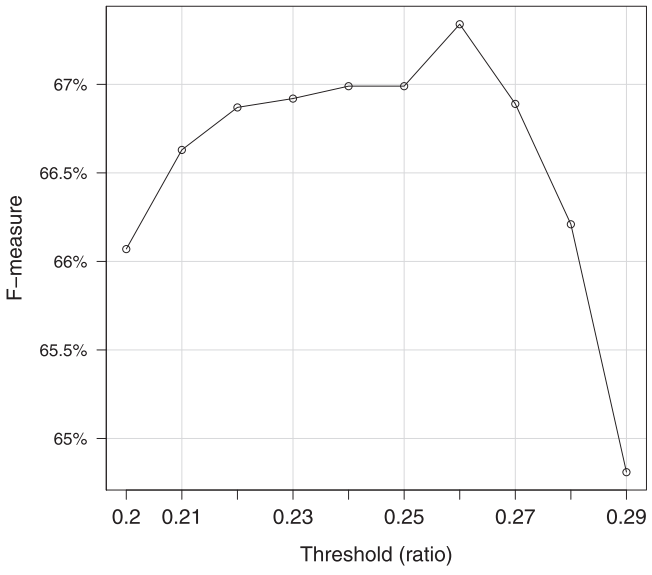


Fig. 7. F-measure by ratio of controversial sections.

To use the sections classifier for further classification, we chose a cut-off value for which its recall in the controversial class was 79% and its F-measure was 68.6%. Fig. 7 shows classification performances based on the ratio of pre-classified sections. We can see that we obtained the best controversy classification for articles having a ratio of controversial sections of 26% or more.

We can also use the classifier based on the EP of sections for classifying the controversy of entire pages. However, if we compare performance data from Fig. 7 with those from Table 1, we can see that the direct use of sentiment measures on entire talk pages is more efficient despite the dilution effect described above. We can use this classifier to locate controversial topics in the article's talk page. By doing so, we can better determine the "essence" of the particular controversy, because many controversial articles consist largely of consensus text, which has been stripped of controversy. Using our EP classifier, we can identify and locate otherwise invisible controversy in the talk page.

The novelty of our approach is the increased granularity of controversy detection in Wikipedia, such as the one with an AUC as high as 0.80 and without any available meta data. We can now detect controversial issues in articles at the sections level of the discussion page.

6. Controversial categories in Wikipedia

Having described our approach for detecting controversy in Wikipedia, in this section we explain how to detect controversial Wikipedia categories.

The Wikipedia category system is a directed cyclic graph, in which articles and categories are nodes and directed edges indicate the parent-child relationship. Every category or article can be a child of many other categories and can have many subcategories. The main Wikipedia category is *Content*, which has seven subcategories. Currently there are over 1.5 M subcategories in Wikipedia. A category may consist of both articles and subordinate categories. Leaf categories in the graph consist of only articles.

While we have a good model for determining if a single article is controversial, we need to aggregate micro scores and determine the controversy of each node (i.e., category) in the Wikipedia category graph, which will make it possible to detect new and potentially controversial articles. For each category, we take all its directly subordinate articles and calculate a mean prediction controversy score. This web content controversy detection system was presented by Jankowski-Lorek and Zieliński (2015).

6.1. Dataset

Currently, the English Wikipedia consists of roughly 5 M articles. We computed all the required features and used the model described in the previous section to determine the controversy score and this score's confidence level for each article. For further research, this dataset is publicly available and can be downloaded from the figshare.com website.[1]

[1] https://figshare.com/articles/Wikipedia_articles_controversy_classification_results_csv/5188462 .

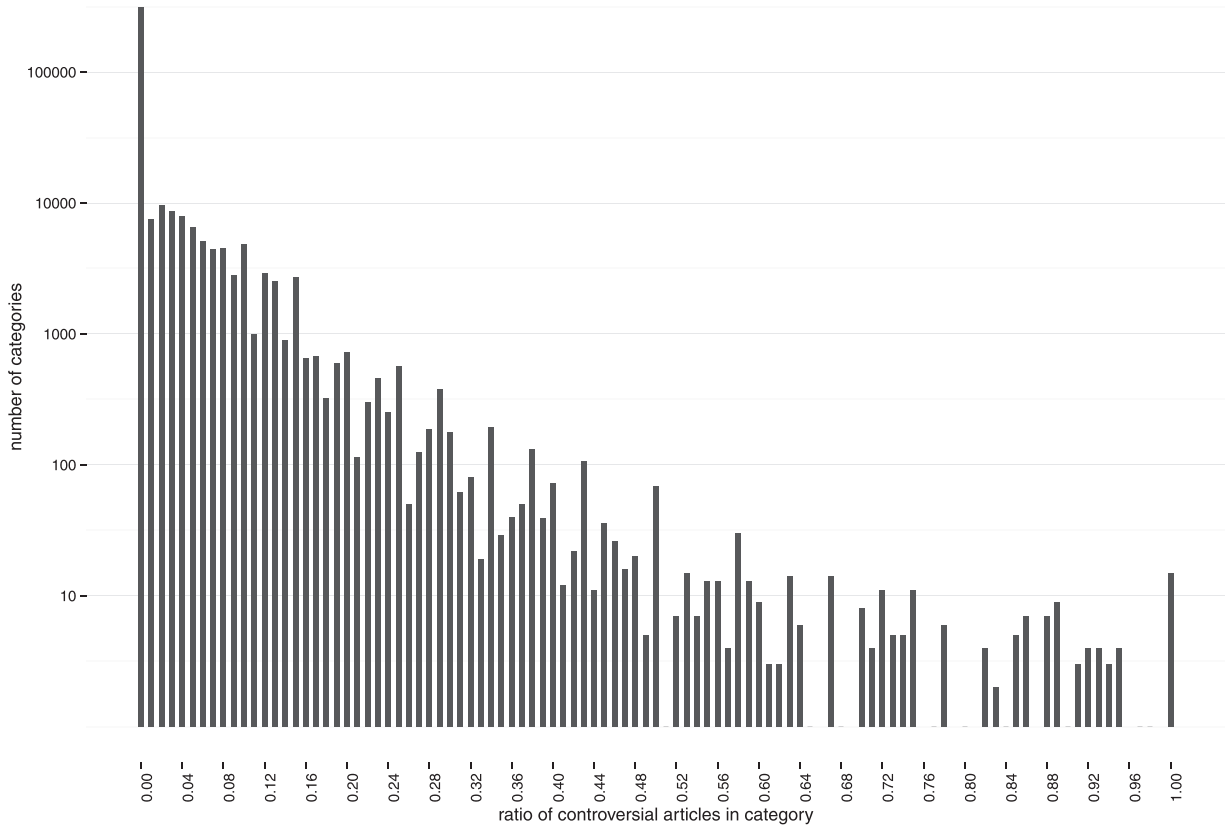


Fig. 8. Histogram of ratio of controversial articles in categories.

Only 0.5% (23,103) of articles are classified as controversial. We consider this to be a plausible result, as the official list of controversial Wikipedia articles contains just 963 entries and in the discussion pages we identified 2153 entries using the controversial article template. This list is not often updated and definitively does not contain all the controversial topics and articles. For each article, we retrieved all the categories to which it is assigned. As described above, an article can belong to several categories and different levels of the Wikipedia category graph. Most articles have fewer than 20 assigned categories but some have more than 200. There are 354,940 categories with only one or two pages assigned. We will discard those categories in future studies as it is not meaningful to aggregate such a small number of articles per category.

6.2. Controversy of Wikipedia categories

Fig. 8shows the number of categories with a given percentage of controversial articles. The trend in distribution is nearly exponential for ratios smaller than 0.5. Above 0.5, i.e., where the number of controversial articles is greater than non-controversial ones, the distribution is irregular with a notable peak at 15 categories with a ratio equal to 1.0. This peak corresponds to categories containing only controversial articles. Only 346 categories have 50% or more controversial articles. It is reasonable to treat these categories as controversial.

Fig. 9shows a histogram of mean confidence level of controversy in Wikipedia categories. For categories with the mean confidence level lower than 0.5 the trend of the distribution of the histogram is nearly logarithmic. This trend is more steady than in Fig. 8. There are 514 categories that have the mean confidence level of controversy of 0.5 or higher. It confirms that the vast majority of Wikipedia content is non-controversial and the probability of finding a random controversial wikipedia category is low. This means that in general, Wikipedia should be considered as not controversial.

The right sides of both histograms in Figs. 8 and 9 (for values on x-axis between 0.5 and 1.0) look less regular in terms of the distribution trend. This may be attributable to a much lower total number of categories in these intervals, preventing any clear trend from appearing.

We tested our method for detecting controversial categories by segregating the article scores (mean confidence level of controversy) based on two approaches. In the first, we manually checked the selected list of categories and in the other we performed a cross validation.

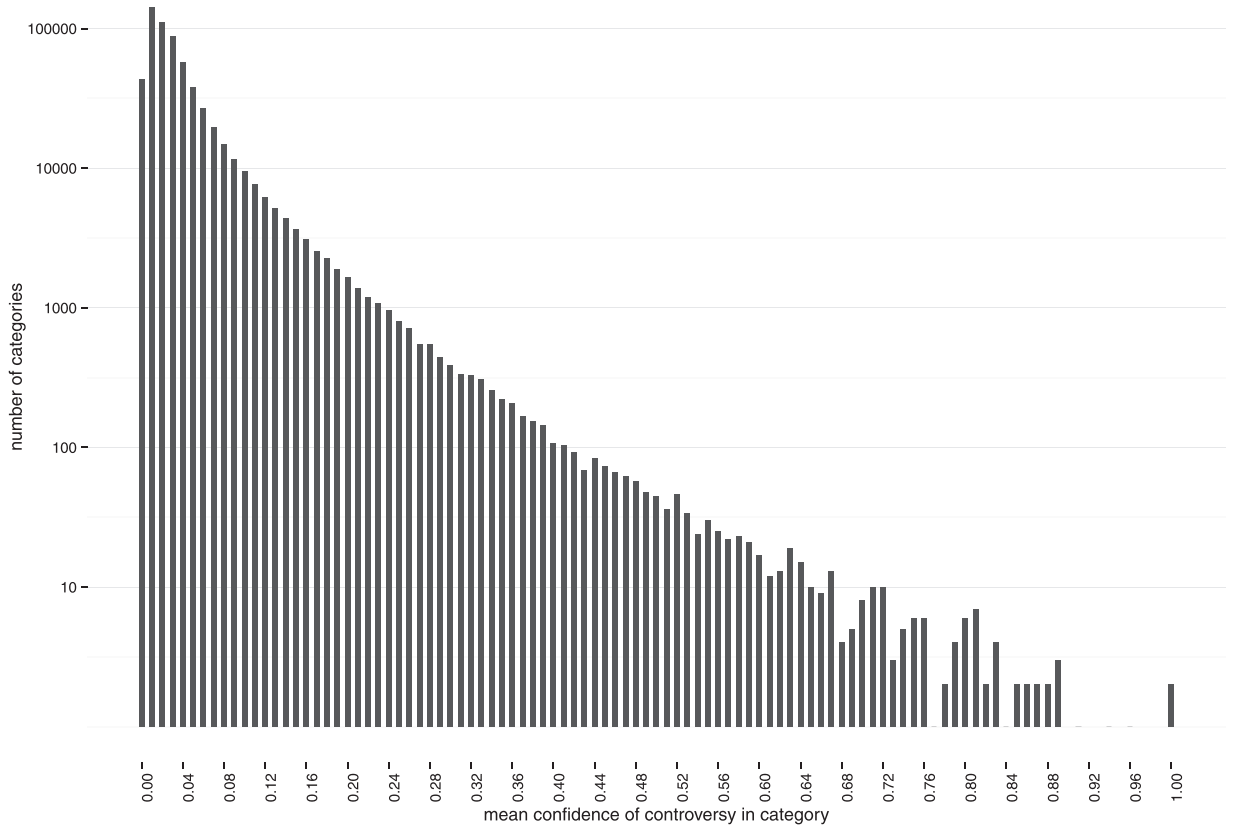


Fig. 9. Histogram of mean confidence level of controversy in categories.

Table 2
Top 20 controversial categories.

Rank	Category	Mean Confidence	# of Articles
1	G8 nations	0.9978125	8
2	G7 nations	0.9975	7
3	G20 nations	0.952368421	19
4	Hindustani-speaking countries and territories	0.935833333	3
5	NUTS 1 statistical regions of the United Kingdom	0.905	3
6	People banned from entering China	0.888333333	3
7	Federal constitutional republics	0.8875	9
8	Slavic countries and territories	0.883076923	13
9	People of the American Enlightenment	0.87375	6
10	Near Eastern countries	0.873333333	9
11	Wars involving Qatar	0.861666667	3
12	Member states of the Union for the Mediterranean	0.860064103	39
13	Member states of NATO	0.855535714	28
14	Member states of the South Asian Association for Regional Cooperation	0.85375	8
15	Middle Eastern countries	0.842916667	18
16	Mormonism	0.841666667	3
17	Northeast Asian countries	0.8325	7
18	Member states of the Council of Europe	0.828928571	14
19	Democratic-Republican Party Presidents of the United States	0.828125	4
20	Western Asian countries	0.827368421	19

6.2.1. Empirical validation

Based on the real-world understanding that some topics are generally known to be controversial, e.g., politics, religion, racism, to name a few, we manually tested the lists of controversial and non-controversial categories.

Table 2contains the top 20 controversial categories based on the mean confidence level of articles. At the top of the list is “G8 nations”, which seems to be the most controversial content-related category in the English Wikipedia. All the top 20 categories are related to politics or religion. Of the top 300 categories, we manually verified 240 as being either controversial or belonging to topics well known to be controversial. In 100 randomly selected categories from the list of all

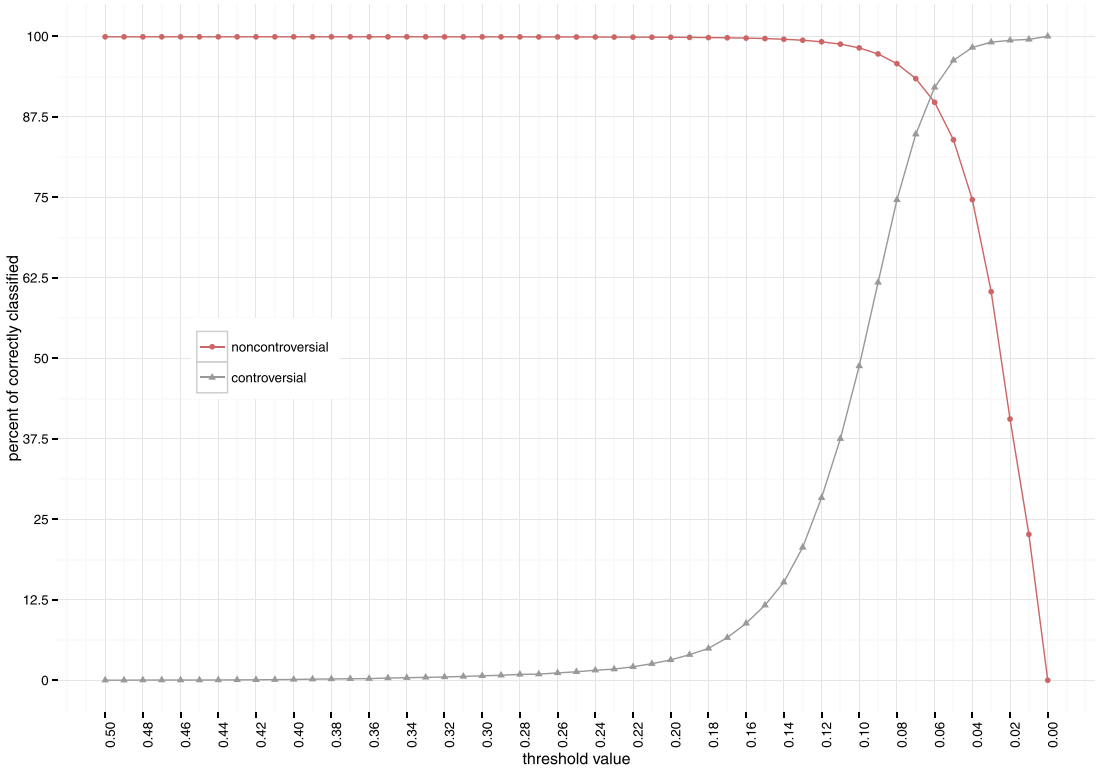


Fig. 10. Percentage of correctly classified controversial and non-controversial articles in correspondence with the chosen threshold value.

controversial categories based on mean confidence, we verified 82% as being controversial. Out of 100 randomly selected non-controversial categories, only 5% were in fact controversial topics (false negatives). This validation result confirms that we can use the mean confidence level of the controversy of articles to find controversial categories.

6.2.2. Cross validation

For the second validation, we randomly split our dataset into training and test subsets at a ratio of 0.7. We used the training subset to calculate the controversy level of categories based on the ratio of controversial to non-controversial articles and the mean controversy confidence level. Then we used these categories to determine which articles in the test subset were controversial, based on the average level of confidence of all the categories to which they were assigned. In the testing dataset, there were 1,399,303 articles, 6926 of which were originally classified as controversial.

Fig. 10shows the percentage of correctly classified controversial and non-controversial articles in relation to the chosen threshold value. An article was assigned as being controversial if the mean confidence controversy level of its categories exceeded this threshold.

We can see that the percentage of correctly classified controversial articles rapidly increases upon lowering the threshold value to below 0.2 and for values up to 0.06, whereas the percentage of correctly classified non-controversial articles holds steady until 0.12 and then begins to slightly decrease and then drops rapidly after reaching the 0.05 threshold value. Based on these results, we chose 0.06 as the best threshold value, with which we can detect 92% of all controversial articles and produce only 10.3% of false positives. The results of our second validation also confirmed that we can use the mean level of confidence in the controversy of articles to identify controversial categories, although in order to correctly detect new controversial articles we must choose an appropriate threshold value.

7. Controversy of Web queries

A Web search engine is the starting point for most information seekers. Users enter search queries in natural language or as a list of keywords. The search engine returns a page containing a list of results (Search Engine Results Page or *SERP*) containing result links sorted by *relevance*. If the query relates to a controversial topic, then the SERP will contain links to pages covering various types of controversial information about that topic. A user will typically select and follow one or only a few links from this list. Jansen and Spink found in Jansen and Spink (2006) that more than 70% users of AltaVista and AllTheWeb search engines viewed only one result page per search so at least they did not follow more links than those from the first result page. More specifically Pass et al. discovered that 67% of all the clicks on SERP fall on the top 3 links on

the first result page (Pass, Chowdhury, & Torgeson, 2006). In case of controversy, an early warning should be issued before any link is followed, because many users never return to the SERP once they have followed a result link.

User's *search intent* is reflected in the *search query*. The mapping of the search intent into the search query is outside of this work. In this research, we focus on the search query and its relation to known controversies.

A search query will be referred to as controversial when it is related to a topic that is known to be controversial. In our approach, we represent this topic by a Wikipedia category. As described in Section 6, we have assigned a controversy score to every category, based on the calculation of the mean score of the controversy classifier (see Section 5.3.1) for all articles belonging to a given category. We shall now compute a similarity function of a query and a category, and use a threshold to determine whether the query is related to the category.

7.1. Queries dataset

Today's search-engine providers consider a database of search terms to be a very valuable and privacy-sensitive asset. Probably for these reasons, the only larger scale dataset of search engine query logs freely available for research purposes is the AOL dataset collected during 2006 and extensively described by Pass et al. (2006). *This collection consists of 21 million web queries, 10M of which are unique*. The query logs were collected from 650k US-based users over a three-month period. This dataset has served as base for many researchers since its publishing in 2006 to recent years (e.g., Shokouhi, 2013; Whiting & Jose, 2014). However, our research does not particularly depend on this specific dataset. In fact, we could use other query logs with our method on equal terms, i.e., as a source of real users' queries.

7.2. Similarity function

Next, we need a function for calculating the similarity between a query and a Wikipedia category, which we define as described below. First we concatenate the texts of all articles belonging to the category and find keywords using the tf-idf vectors in the corpus of all Wikipedia articles. Then, we match the top-N keywords against the words in the query and calculate the aggregated tf-idf values for M matching words and N keywords, respectively.

$$Q_{Sim}(q,c) = \frac{\sum_{i=1}^{M} Q_i}{\sum_{i=1}^{N} C_i} \tag{3}$$

Q_i ... tf-idf value of *i-th* matching word of query q
C_i ... tf-idf value of top-*i-th* keyword of category c
$M \leq N$

Q_{Sim} takes values from the range [0, 1]. A value of 0 means there is no similarity, while 1 indicates *perfect* similarity. A score of 1 can be obtained only if all N keywords in the category are present in the query. Because most queries are shorter than four words, four is a good value for N. Higher numbers for N are also good, but the Formula (3) result will yield a smaller number. Since we use this number for ranking, the exact value of N does not really matter.

When a user enters a query in natural language, the number of meaningful words in the query can easily exceed four. In such cases, more than N keywords should be used in the similarity formula. If the match of the top N keywords is not *perfect*, then we can extend the outlook to additional top-L keywords. If the number of additional matching keywords is K then their tf-idf values are summed to R:

$$R = \sum_{i=1}^{K} Q_i \tag{4}$$

Then

$$Q_{Sim}(q,c) = \frac{\sum_{i=1}^{M} Q_i + R}{\sum_{i=1}^{N} C_i + R} \tag{5}$$

In Formula (5), keywords 1..N are *mandatory*, i.e., the absence of a *mandatory* keyword in the query decreases the similarity score, and keywords N+1..L are *optional*, i.e., the presence of an *optional* keyword in the query increases the score obtained by using only the *mandatory* keywords. The range of the function Q_{Sim} is [0, 1].

7.3. Early warning algorithm

Having defined the Q_{Sim} function in Section 7.2, we can now measure the similarity of any search term to any concept represented by a Wikipedia category, e.g., for every query we can calculate its similarity scores for all categories. However, as we already know the controversy score for each Wikipedia category, and we are looking only for controversy, for practical reasons we limit this task to categories with a controversy score exceeding some arbitrary threshold value. We define a controversy sub-score on the level query-category by the following formula:

$$Sub_{Cont}(q,c) = Q_{Sim}(q,c) \cdot score(c). \tag{6}$$

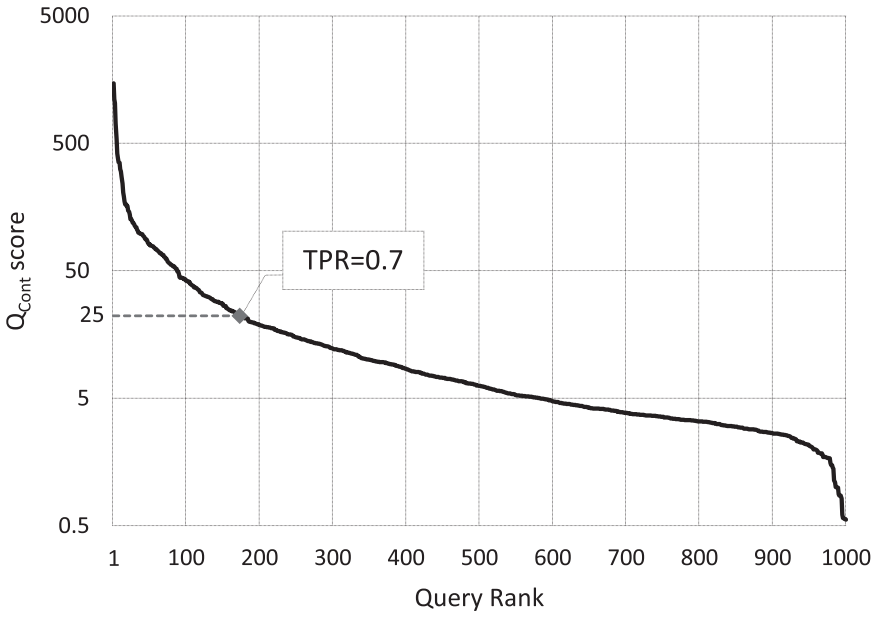


Fig. 11. Q_{Cont} score for top 1000 controversial queries.

For every query, we can have as many Sub_{Cont} scores as we have categories exceeding the threshold value. Again, because we are looking only for controversy, we consider only those pairs (query, category) for which their Q_{Sim} exceeds some arbitrary threshold. As a result, we can have 0 to j Sub_{Cont} scores for each query. Then, we can introduce the controversy score of a query as follows:

$$Q_{Cont}(q) = (1-\alpha)\cdot \max(Sub_{Cont}(q, c_i)) + \alpha \cdot \sum_{i=1}^{j}(Sub_{Cont}(q, c_i)). \tag{7}$$

By manipulating the value of parameter α between 0.0 and 1.0, we can balance the impact of the highest match against the aggregated value of all matches for a given query.

We calculated Q_{Cont} score for top 1000 controversial queries and ranked them by the decreasing scores, i.e., the highest score first. Fig. 11 shows the distribution of Q_{Cont} score by the decreasing rank for the top 1000 controversial queries. A rapid fall of Q_{Cont} score can be observed on the logarithmic y-axis with the decreasing rank on x-axis. This indicates that a binary classification by a cut-off value of Q_{Cont} may perform quite well.

7.4. Empirical validation

We have manually validated the top-ranked results of the *Early Warning Algorithm* for the existence of controversy. The validation was performed by five annotators who were asked to evaluate controversy of topics that they associated with the query, using their own knowledge regarding these topics. Annotators were asked to look at the topics from the perspective of their communities (annotators were Polish, but fluent in the English language and familiar with contemporary global culture). If annotators had no knowledge regarding topics associated with the query, they were allowed to use Google searches.

Annotators assigned a binary evaluation of query controversy. If controversy was discovered, they were also asked to briefly justify their evaluation by describing the topic that was related to the query and caused lots of controversy. The final decision regarding a query's association with controversial topics was made based on majority decision of five annotators. Table 3 lists the queries and the related topics that were the reason for the evaluation of the query as controversial. Significantly, within the top 25 highest-ranked queries using the proposed controversy ranking, the annotators found only two false positives (true positive rate of 92%). All remaining queries were evaluated as controversial.

On the other end of the ranked result list there are queries that are least likely to refer to controversy. Within 25 such queries shown in Table 4 only three false negatives were found (true negative rate of 88%). Some of the examples of the top 25 queries related to controversial topics clearly illustrate the potential of an "early warning" about controversy. Consider the query "selena" (position 8 in the controversy ranking). A superficial Google search (looking just at the top 10 search results) allows the user to find information about the pop-star Selena Gomez.[2] However, without digging deeper in the

[2] https://www.selenagomez.com/ .

Table 3
Top controversial queries found in dataset.

	Query	Controversy	Notes
1	puerto rico	YES	migration related issues in USA
2	continental airlines	YES	past airplane crashes
3	harry potter	YES	occult and satanic subtexts in novel for children
4	vietnam war	YES	strong US-American opposition against the war
5	abortion	YES	moral issue brought up on political agenda in many countries
6	thomas jefferson	YES	relationship of US president with his mixed-race slave Sally Hemings
7	fidel castro	YES	popular champion or cruel dictator
8	selena	YES	disrespectful photo in mosque
9	microsoft windows	YES	monopolist business practices, security issues
10	rush limbaugh	YES	anti-immigration statements
11	korea	YES	Korean history textbook in South Korea, int'l criticism of North Korean politics
12	bill clinton	YES	Lewinsky scandal
13	pope john paul ii	YES	lack of response to abuse of children in church
14	star trek	NO	Not controversial. Sexuality in original "Star Trek" series.
15	sierra leone	YES	Ebola lockdown policy
16	osama bin laden	YES	death conspiracy theories
17	hillary clinton	YES	private use of governmental email
18	mormons	YES	book of Mormon origins, polygamy
19	mario games	YES	occult, animal cruelty, transgenderism
20	nigger	YES	racist insult
21	amelia earhart	YES	death conspiracy theories
22	airline flights	NO	Not controversial. Relationship to terrorism.
23	victoria beckham	YES	Use of skinny models in fashion shows
24	haile selassie	YES	dictator and religious leader overthrown by Marxist coup
25	space shuttle challenger	YES	Criticism of the Space Shuttle program in US

Table 4
Example least controversial queries in dataset.

	Query	Controversy	Notes
1	b-box	NO	baby essentials manufacturer
2	boxes and more	NO	packaging products supplier
3	how to box	NO	
4	republican candidates	NO	
5	where is poitiers	NO	
6	what is manga	NO	
7	nixon election	YES	
8	donate computer	NO	
9	yankees	NO	
10	3m corp	NO	
11	center city	NO	
12	development of the brain	NO	
13	what to do in a flood	NO	
14	smoking and its effects	YES	Smoking of marihuana
15	what is ba	NO	
16	avocado what is in it	NO	
17	star symbols	YES	Occult, political controversies
18	more than coffee	NO	cafe chain
19	j.a.c. redford	NO	
20	art and music	NO	
21	concrete how to	NO	
22	how to paint furniture	NO	
23	messina italy	NO	
24	city of peace	YES	Jerusalem, Baghdad, Auschwitz
25	wild stallions	NO	

Google search results, or using the query "selena mosque", the user would not learn about a controversy about Selena in Muslim countries (referring to allegedly inappropriate behavior in a mosque).

Another example is the query "mario games" (position 19 in the controversy ranking). Again, the Google top 10 search results will not reveal information about controversy. However, the query "mario games cruel" can retrieve a Web page of the "mariowiki"[3] that contains a list of controversies related to the game, including concerns about animal cruelty, transgenderism and the occult. One of the false positives in the top 25 is the query "Star Trek". Interestingly, it seems that in the 1960s and 1970s, when the original "Star Trek" series was running, it may have generated significant controversy in

[3] https://www.mariowiki.com/List_of_Mario-related_controversies .

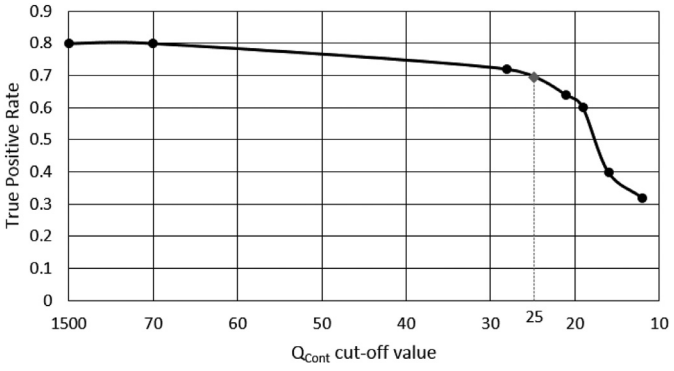


Fig. 12. TPR of binary classification by cut-off value.

the United States because of references or suggestions of homosexuality, polyamorous relationships etc.[4] It seems that since then this controversy had been reduced, but its traces still remain on the Web and Wikipedia. The relation between controversy and time is an interesting issue that we plan to investigate more in the future. Like almost everything, controversy too tends to fluctuate and can occur over certain periods only. The proposed approach mainly detects controversies existing currently or in the recent past.

Some of the false negatives in the bottom 25 of the query list demonstrate difficulties of automatic controversy evaluation. For example, the query "star symbols" was classified as non-controversial. However, star symbols include the "Star of David" (symbol of Israel which sometimes appears in anti-semitic controversies) and the pentagram (used in occult). Such cultural knowledge is difficult to teach to an algorithm. Another false negative, the query "smoking and its effects" can refer to tobacco (in this context, it may no longer be controversial), but it can also refer to marijuana. The algorithm cannot simply guess this association, and the wording of the query was apparently understood by the algorithm as referring to smoking of tobacco.

Having ranked the top 1000 results (see Fig. 11) we now need a cut-off value for the binary classification: if $Q_{Cont} >$ cut-off, then the query is classified controversial, otherwise it is classified not controversial. In order to choose the cut-off value we first had to estimate the performance of such a classification, for which we had to use human annotation again. We divided the top 250 ranked queries in 10 bins, 25 queries each. Each bin spreads over a decreasing range of Q_{Cont}. The controversy for each query was manually annotated in bins 1, 2, 5, 6, 7, 8 and 10. Then we calculated True Positive Rate (TPR) for each of the annotated bins and presented it graphically in Fig. 12. This plot lets us estimate that with an example cut-off value of Q_{Cont} score of 25 we obtain TPR of 0.7. The position of the same example cut-off point is also shown on the ranking plot, see Fig. 11.

7.5. Robustness to manipulation

One concern about the proposed approach relates to adversarial manipulation of controversy computation. It could be possible that some entities try to impact the results to either make certain Wikipedia articles seem less controversial or, in contrast, to "render" the non-controversial ones such that they become controversial. However, this would require manipulating data on several levels. In particular, not only the perceptions of trustworthiness given by Wikipedia visitors should be impacted but also the sentiment levels of Wikipedia talk pages, which record communication between the editors, as well as the numbers of edits, revisions, unique editors, etc. This is rather difficult, especially, considering that Wikipedia categories may contain large number of individual pages. To summarize, the use of several independent variables by the controversy classifiers, as well as the aggregation of article controversy scores into categories makes our method quite robust to adversarial manipulation.

8. Conclusions and future work

Detecting controversy is a challenge that reaches far beyond natural language processing algorithms and related research. It requires a holistic approach and different tools drawn from Web science and social informatics. The results presented here demonstrate that our formal model, introduced in Section 3, correctly identifies potential problems associated with predicting controversy. The proposed formal model of controversy might also be useful as a basis for further study and introduces a common denominator for use by very different research perspectives.

Wikipedia articles provide a good starting point for learning about controversy due to the vast amount of meta information that can be used for predicting and explaining sources of disagreement. As shown in Table 1 in Section 5, combining

[4] https://en.wikipedia.org/wiki/Sexuality_in_Star_Trek .

our method with previously proposed approaches boosts the F-measure to 84.10% and constitutes a good basis for further research. We used the predicted controversy of Wikipedia articles to calculate the list of the most controversial categories. This list might be useful for the selection of topics for more careful monitoring. This is a subject important to the Wikipedia community, because Wikipedia aims to realize a neutral point of view and excludes excessive controversial views.

In Section 7, we showed that evaluating controversy of Wikipedia content may also be used for identifying controversial queries in search engines. We tested our proposed algorithm on real queries submitted by people. Our manually conducted validation shows that it works quite well for most controversial queries while not raising too many false alarms. We emphasize that a query with a controversy warning should not be considered as negative or harmful, but rather the warning should be seen as a useful recommendation for users to be careful when studying any search results. Moreover, because of the heavy personalization and the filter bubble effect on search results, a particular user might only retrieve one point of view, so a controversy warning indicates that some important information may be lacking in the search results.

Controversies are inevitable. Therefore, further research on controversy on the Internet is needed to better understand its dynamics. New algorithms for detecting and mitigating controversies will definitely be developed in the foreseeable future. In our future work, we plan to verify whether the use of content and meta information from Q&A websites might help to identify controversy with respect to rapidly changing topics or topics that are beyond the coverage of Wikipedia. We also plan to include time and community dynamics in our formal model of controversy.

The remainder of this section discusses the limitations of our study, possible social ramifications of alerting Web users to query controversy, and future work.

8.1. Study limitations

Our study follows an empirical research methodology. In this section, we describe the limitations of our research methods, and possible ways of addressing them.

8.1.1. AOL Query Log

AOL Query Log was published in 2006, hence, it might seem outdated. We have utilized AOL Query Log due to its availability and relatively large size. To the best of our knowledge, no other publicly available dataset of Web search queries has a similar size (about 10 million unique queries). While the set of queries is over 10 years old and the Web search behavior of users may have changed since then, our estimations of controversy of topics related to the queries have been performed on the current versions of Wikipedia articles. Similarly, the evaluation described in Section 7.4 was performed under current circumstances (i.e., as if the queries were issued currently). Hence, there is no problem of time difference or the lack of synchronization between the knowledge base used for computing the results and the time of evaluation.

8.1.2. Web users vs. Wikipedia users

Our research has made the assumption that controversy evaluation based on Wikipedia will be useful for all Web users. This is the same as assuming that the community of Wikipedia users will provide a useful controversy evaluation for the community of all Web users. Even if we restrict these communities by language (English-language Wikipedia and English-language Web), there may be Web users who do not use Wikipedia. However, according to a recent study,[5] despite a drop in Google search rankings of Wikipedia pages, over 50% of search queries have a Wikipedia page in the top 10 of Google search results. Wikipedia has about 18 billion views per month[6] (as compared to about 100 billion Google searches monthly). The number of users is harder to estimate, but the number of unique devices that access the English-language Wikipedia daily is about 60 million.[7] These statistics show that while the community of Wikipedia users is smaller than the community of all Web users (for the English language), it should be sufficiently large to be treated as a large sample (at least on the order of tens of millions in size). Nevertheless, little is known about the community of Web users that does not use Wikipedia, and more research of this subject is required.

What is more significant is that according to Wikipedia policy, Wikipedia only excludes content that is not encyclopedic. Wikipedia is adding 20,000 new articles each month. It reacts very quickly to current newsworthy events, such as the death of Michael Jackson or the Fukushima catastrophe. For these reasons, it is likely that controversies that occur in the community of Web users will eventually (rather sooner than later) be reflected in the Wikipedia.

8.1.3. Cultural biases

Naturally, culture and background of a person impacts the way in which she or he judges and perceives controversy. Conformist societies such as Japanese society may have different attitude and propensity than individualistic societies such as the USA towards evaluating and measuring controversy. The proposed method can be adapted to different cultures by utilizing different language versions of Wikipedia (e.g., Japanese Wikipedia). We believe that this could partially mitigate the problem of cultural impact and bias.

---

[5] https://www.stonetemple.com/google-still-loves-wikipedia-more-than-its-own-properties/ , September 23, 2015.
[6] http://www.pewresearch.org/fact-tank/2016/01/14/wikipedia-at-15/ .
[7] https://analytics.wikimedia.org/dashboards/reportcard/#daily-unique-devices , July, 4th, 2017.

8.2. Socio-technical Ramifications

Alerting users about the controversy levels of visited Wikipedia articles or, at a finer granularity, about shown paragraphs, should help them decide if certain information can be trusted or not. In case of controversial contents, controversy alerts should give them a chance to perform more research and investigation. Automatic labeling of Wikipedia articles and categories as for their controversy levels would then improve the experience of using Wikipedia and decrease the uncertainty regarding the correctness of information encountered within Wikipedia. While the main objective of the current approach is to detect the controversy, future work will investigate methods for explaining the reasons behind the controversy. This should provide further benefits to readers since showing scores alone may be just seen as unreliable or ad hoc by users. Without solid proof, readers may have difficulty in accepting and understanding the output controversy scores.

As far as tagging the search queries with the controversy values is concerned, the proposed technique could be incorporated into existing search engines. For example, when typing the query users could be provided with the pre-computed controversy levels of concepts underlying the query. This would allow them to distinguish between controversial concepts and non-controversial ones, thus enabling a more informed search.

The last issue to mention is the effect of false positives on search or analysis of topics. Naturally, no system is always perfect and, thus, in some cases wrong labels may be assigned to either Wikipedia contents or to search queries. We think that this should not happen to clearly controversial or clearly non-controversial information, but it likely occurs to rather "gray" contents. We also believe that outputting the controversy reasons in future should decrease chances of users being misinformed.

8.3. Future work: attribute-based search approach

A given query, especially, one denoting some general concept or complex entity, may have some of its related aspects controversial, while others may not be controversial at all. The controversy score for such a query should be seen as the sum of partial controversy scores assigned to its aspects or related topics. Let us take the example of Alexandre Dumas, a topic deemed non-controversial in Section 3.1, which can easily become controversial – his recent portrayal in a French biopic by Gerard Depardieu was very debated and much maligned. Indeed if a search query relates to Alexandre Dumas, then its closest matching Wikipedia category will be Alexandre Dumas. This category currently contains 12 articles, one of them being about the French film "Dumas" of 2010 (original title: L'Autre Dumas). Hence the controversy about the film Dumas will contribute to the controversy score of the category Alexandre Dumas with a weight of 1/12.

A possible approach towards improving the current proposal would be to detect attributes of entities contained in queries and then to investigate their individual controversy levels. Such a method would necessarily need to utilize semantic data about the entities (e.g., using DBpedia[8] or Yago[9]) or would need to rely on applying effective natural language processing technologies for successful extraction of entity aspects and properties. In future we plan to investigate semantic approach for controversy detection as well as, as mentioned before, explicitly incorporate temporal aspects into the recognition model.

Acknowledgment

This work is supported by Polish National Science Centre grant 2015/19/B/ST6/03179.

References

Biltereyst, D. (2004). Media audiences and the game of controversy. *Journal of Media Practice, 5*(1), 7–24.

Biuk-Aghai, R. P., Pang, C.-I., & Si, Y.-W. (2014). Visualizing large-scale human collaboration in Wikipedia. *Future Generation Computer Systems, 31*, 120–133. doi:10.1016/j.future.2013.04.001.

Borra, E., Weltevrede, E., Ciuccarelli, P., Kaltenbrunner, A., Laniado, D., Magni, G., … Venturini, T. (2015). Societal controversies in Wikipedia articles. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. In *CHI '15* (pp. 193–196). New York, NY, USA: ACM. doi:10.1145/2702123.2702436.

Borzymek, P., Sydow, M., & Wierzbicki, A. (2009). Enriching trust prediction model in social network with user rating similarity. In *Computational aspects of social networks, 2009. CASON '09. International conference* (pp. 40–47). IEEE Computer Society. doi:10.1109/CASoN.2009.30.

Buriol, L. S., Castillo, C., Donato, D., Leonardi, S., & Millozzi, S. (2006). Temporal analysis of the Wikigraph. In *Proceedings of Web intelligence, Hong Kong, 2006* (pp. 45–51).

Chandy, R. (2008). Wikiganda: Identifying propaganda through text analysis. *Caltech Undergraduate Research Journal. Winter, 2009*, 6–11.

Coser, L. (1956). The functions of social conflict. *Free press paperback*. Free Press.

De Clercq, O., Hertling, S., Hoste, V., Ponzetto, S. P., Paulheim, H., et al. (2014). Identifying disputed topics in the news. *Linked Data for Knowledge Discovery (LD4KD), CEUR*, 37–48.

Dori-Hacohen, S. (2015). Controversy detection and stance analysis. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. In *SIGIR '15*. New York, NY, USA: ACM. doi:10.1145/2766462.2767844. 1057–1057

Dori-Hacohen, S., & Allan, J. (2013). Detecting controversy on the Web. In *Proceedings of the 22nd ACM international conference on conference on information and knowledge management*. In *CIKM '13* (pp. 1845–1848). New York, NY, USA: ACM. doi:10.1145/2505515.2507877.

Dori-Hacohen, S., & Allan, J. (2015). Automated controversy detection on the Web. In A. Hanbury, G. Kazai, A. Rauber, & N. Fuhr (Eds.), *Advances in information retrieval*. In *Lecture Notes in Computer Science: 9022* (pp. 423–434). Springer International Publishing. doi:10.1007/978-3-319-16354-3_46.

[8] http://wiki.dbpedia.org .

[9] http://www.yago-knowledge.org/ .

Dori-Hacohen, S., Jensen, D., & Allan, J. (2016). Controversy detection in Wikipedia using collective classification. In *Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval*. In *SIGIR '16* (pp. 797–800). New York, NY, USA: ACM. doi:10.1145/2911451.2914745.

Dori-Hacohen, S., Yom-Tov, E., & Allan, J. (2015). Navigating controversy as a complex search task. In *Proceedings of the first international workshop on supporting complex search tasks, ECIR, 2015: 1338*.

Van der Eijk, C. (2001). Measuring agreement in ordered rating scales. *Quality and Quantity, 35*(3), 325–341.

Ennals, R., Byler, D., Agosta, J. M., & Rosario, B. (2010a). What is disputed on the Web? In *Proceedings of the 4th workshop on information credibility* In *WICOW '10* (pp. 67–74). New York, NY, USA: ACM. doi:10.1145/1772938.1772952.

Ennals, R., Trushkowsky, B., & Agosta, J. M. (2010b). Highlighting disputed claims on the Web. In *Proceedings of the 19th international conference on World Wide Web*. In *WWW '10* (pp. 341–350). New York, NY, USA: ACM. doi:10.1145/1772690.1772726.

Feldman, R., Rosenfeld, B., Bar-Haim, R., & Fresko, M. (2011). The stock sonar - sentiment analysis of stocks based on a hybrid approach.. *Twenty-third IAAI conference*. Association for the Advancement of Artificial Intelligence.

Ferschke, O., Gurevych, I., & Chebotar, Y. (2012). Behind the article: Recognizing dialog acts in Wikipedia talk pages. In *Proceedings of the 13th conference of the European chapter of the association for computational linguistics*. In *EACL '12* (pp. 777–786). Stroudsburg, PA, USA: Association for Computational Linguistics.

Garimella, K., Mathioudakis, M., Morales, G. D. F., & Gionis, A. (2016). Exploring controversy in Twitter. In *Proceedings of the 19th ACM conference on computer supported cooperative work and social computing companion*. In *CSCW '16 Companion* (pp. 33–36). New York, NY, USA: ACM. doi:10.1145/2818052.2874318.

Garimella, K., Morales, G. D. F., Gionis, A., & Mathioudakis, M. (2015). Quantifying controversy in social media. arXiv:1507.05224,.

Gurr, T. R., & Duvall, R. (1976). Introduction to a formal theory of political conflict. *The Uses of Controversy in Sociology*, 139.

Hajian, B., & White, T. (2011). Measuring semantic similarity using a multi-tree model. In *CEUR workshop proceedings 2011: 756* (p. 7). Citeseer.

Han, M. S. (2013). Semantic information retrieval based on Wikipedia taxonomy. *International Journal of Computer Applications Technology and Research, 2*(1), 77–80.

Jamieson, S., et al. (2004). Likert scales: How to (ab) use them. *Medical Education, 38*(12), 1217–1218.

Jang, M., & Allan, J. (2016). Improving automated controversy detection on the Web. In *Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval*. In *SIGIR '16* (pp. 865–868). New York, NY, USA: ACM. doi:10.1145/2911451.2914764.

Jang, M., Foley, J., Dori-Hacohen, S., & Allan, J. (2016). Probabilistic approaches to controversy detection. In *Proceedings of the 25th ACM international on conference on information and knowledge management*. In *CIKM '16* (pp. 2069–2072). New York, NY, USA: ACM. doi:10.1145/2983323.2983911.

Jankowski-Lorek, M., Nielek, R., Wierzbicki, A., & Zieliński, K. (2014). Predicting controversy of Wikipedia articles using the article feedback tool. In *Proceedings of the 2014 international conference on social computing*. In *SocialCom '14* (pp. 22:1–22:7). New York, NY, USA: ACM. doi:10.1145/2639968.2640074.

Jankowski-Lorek, M., & Zieliński, K. (2015). Document controversy classification based on the Wikipedia category structure. *Computer Science, 16*(2), 185–198.

Jansen, B. J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management, 42*(1), 248–263. doi:10.1016/j.ipm.2004.10.007.

Kaptein, R., Koolen, M., & Kamps, J. (2009). Using Wikipedia categories for ad hoc search. In *Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval*. In *SIGIR '09* (pp. 824–825). New York, NY, USA: ACM. doi:10.1145/1571941.1572147.

Kawahara, D., Inui, K., & Kurohashi, S. (2010). Identifying contradictory and contrastive relations between statements to outline Web information on a given topic. In *Proceedings of the 23rd international conference on computational linguistics: Posters*. In *COLING '10* (pp. 534–542). Stroudsburg, PA, USA: Association for Computational Linguistics.

Kąkol, M., Jankowski-Lorek, M., Abramczuk, K., Wierzbicki, A., & Catasta, M. (2013). On the subjectivity and bias of Web content credibility evaluations. In *Proceedings of the 22nd international conference on World Wide Web*. In *WWW '13 Companion* (pp. 1131–1136). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee.

Kittur, A., Chi, E. H., & Suh, B. (2009). What's in Wikipedia?: Mapping topics and conflict using socially annotated category structure. In *Proceedings of the SIGCHI conference on human factors in computing systems*. In *CHI '09* (pp. 1509–1512). New York, NY, USA: ACM. doi:10.1145/1518701.1518930.

Kittur, A., Suh, B., Pendleton, B. A., & Chi, E. H. (2007). He says, she says: Conflict and coordination in Wikipedia. In *Proceedings of the SIGCHI conference on human factors in computing systems*. In *CHI '07* (pp. 453–462). New York, NY, USA: ACM. doi:10.1145/1240624.1240698.

Koutra, D., Bennett, P. N., & Horvitz, E. (2015). Events and controversies: Influences of a shocking news event on information seeking. In *Proceedings of the 24th international conference on World Wide Web*. In *WWW '15* (pp. 614–624). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. doi:10.1145/2736277.2741099.

Laniado, D., Kaltenbrunner, A., Castillo, C., & Morell, M. F. (2012). Emotions and dialogue in a peer-production community: The case of Wikipedia. In *Proceedings of the eighth annual international symposium on Wikis and open collaboration*. In *WikiSym '12* (pp. 9:1–9:10). New York, NY, USA: ACM. doi:10.1145/2462932.2462944.

Lauw, H. W., Lim, E.-P., & Wang, K. (2006). Bias and controversy: Beyond the statistical deviation. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, 2006* (pp. 625–630). ACM.

Lauw, H. W., Lim, E.-P., & Wang, K. (2008). Bias and controversy in evaluation systems. *IEEE Transactions on Knowledge and Data Engineering, 20*(11), 1490–1504. doi:10.1109/TKDE.2008.77.

Leik, R. K. (1966). A measure of ordinal consensus. *The Pacific Sociological Review, Vol. 9*, 85–90.

Lo, D., Surian, D., Prasetyo, P. K., Zhang, K., & Ee-Peng, L. (2013). Mining direct antagonistic communities in signed social networks. *Information Processing & Management, 49*(4), 773–791.

Medelyan, O., Milne, D., Legg, C., & Witten, I. H. (2009). Mining meaning from Wikipedia. *International Journal of Human-Computer Studies, 67*(9), 716–754. doi:10.1016/j.ijhcs.2009.05.004.

Mehdi, M., Okoli, C., Mesgari, M., Nielsen, F. A., & Lanamaki, A. (2017). Excavating the mother lode of human-generated text: A systematic review of research that uses the Wikipedia corpus. *Information Processing & Management, 53*(2), 505–529.

Mejova, Y., Zhang, A. X., Diakopoulos, N., & Castillo, C. (2014). Controversy and sentiment in online news. *Computation+journalism symposium arXiv:1409.8152*.

Milne, D., & Witten, I. H. (2008). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceeding of AAAI workshop on Wikipedia and artificial intelligence: an evolving synergy, AAAI Press, Chicago, USA, 2008* (pp. 25–30).

Nastase, V., & Strube, M. (2008). Decoding Wikipedia categories for knowledge acquisition. In *Proceedings of the 23rd national conference on artificial intelligence - volume 2*. In *AAAI'08* (pp. 1219–1224). AAAI Press.

Nguyen, T. T., Hui, P.-M., Harper, F. M., Terveen, L., & Konstan, J. A. (2014). Exploring the filter bubble: The effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World Wide Web*. In *WWW '14* (pp. 677–686). New York, NY, USA: ACM. doi:10.1145/2566486.2568012.

Nielek, R., Wawer, A., & Wierzbicki, A. (2010). Spiral of hatred: Social effects in Internet auctions. Between informativity and emotion. *Electronic Commerce Research, 10*, 313–330.

Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F. A., & Lanamaki, A. (2012). The people's encyclopedia under the gaze of the sages: A systematic review of scholarly research on Wikipedia. *Social Science Research Network*. (p. http://papers.ssrn.com/abstract=2021326)

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin UK.

Pass, G., Chowdhury, A., & Torgeson, C. (2006). A picture of search. In *Proceedings of the 1st international conference on scalable information systems*. In *number 1 in InfoScale '06*. New York, NY, USA: ACM. doi:10.1145/1146847.1146848.

Pennacchiotti, M., & Popescu, A.-M. (2010). Detecting controversies in Twitter: A first study. In *Proceedings of the NAACL HLT 2010 workshop on computational linguistics in a world of social media*. In *WSA '10* (pp. 31–32). Stroudsburg, PA, USA: Association for Computational Linguistics.

Rad, H. S., & Barbosa, D. (2012). Identifying controversial articles in Wikipedia: A comparative study. In *Proceedings of the eighth annual international symposium on Wikis and open collaboration*. In *WikiSym '12* (pp. 7:1–7:10). New York, NY, USA: ACM. doi:10.1145/2462932.2462942.

Rafalak, M., Deja, D., Wierzbicki, A., Nielek, R., & Kakol, M. (2016). Web content classification using distributions of subjective quality evaluations. *ACM Transactions on the Web, 10*(4), 21:1–21:30. doi:10.1145/2994132.

Sang, E. T. K., & Bos, J. (2012). Predicting the 2011 Dutch senate election results with Twitter. In *Proceedings of the workshop on semantic analysis in social media* (pp. 53–60). Stroudsburg, PA, USA: Association for Computational Linguistics.

Schonhofen, P. (2006). Identifying document topics using the Wikipedia category network. In *Proceedings of the 2006 IEEE/WIC/ACM international conference on Web intelligence*. In *WI '06* (pp. 456–462). Washington, DC, USA: IEEE Computer Society. doi:10.1109/WI.2006.92.

Sepehri-Rad, H., & Barbosa, D. (2015). Identifying controversial Wikipedia articles using editor collaboration networks. *ACM Transactions on Intelligent Systems and Technology, 6*(1), 5:1–5:24. doi:10.1145/2630075.

Shokouhi, M. (2013). Learning to personalize query auto-completion. In *Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval*. In *SIGIR '13* (pp. 103–112). New York, NY, USA: ACM. doi:10.1145/2484028.2484076.

Smith, L. M., Zhu, L., Lerman, K., & Kozareva, Z. (2013). The role of social media in the discussion of controversial topics. In *Proceedings of the 2013 international conference on social computing*. In *SOCIALCOM '13* (pp. 236–243). Washington, DC, USA: IEEE Computer Society. doi:10.1109/SocialCom.2013.41.

Stvilia, B., Twidale, M. B., Smith, L. C., & Gasser, L. (2005). Assessing information quality of a community-based encyclopedia. In *Proceedings of the international conference on information quality 2005* (pp. 442–454).

Sumi, R., Yasseri, T., Rung, A., Kornai, A., & Kertész, J. (2011a). Characterization and prediction of Wikipedia edit wars. In *Proceedings of the ACM WebSci 11*. The Web Science Trust.

Sumi, R., Yasseri, T., Rung, A., Kornai, A., & Kertesz, J. (2011b). Edit wars in Wikipedia. In *IEEE third international conference on social computing (SocialCom 2011)* (pp. 724–727). doi:10.1109/PASSAT/SocialCom.2011.47.

Tastle, W. J., & Wierman, M. J. (2007). Consensus and dissention: A measure of ordinal dispersion. *International Journal of Approximate Reasoning, 45*(3), 531–545.

Tsytsarau, M., Palpanas, T., & Denecke, K. (2010). Scalable discovery of contradictions on the Web. In *Proceedings of the 19th international conference on World Wide Web*. In *WWW '10* (pp. 1195–1196). New York, NY, USA: ACM. doi:10.1145/1772690.1772871.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2011). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review, 29*(4), 402–418. doi:10.1177/0894439310386557.

Turek, P., Wierzbicki, A., Nielek, R., Hupa, A., & Datta, A. (2010). Learning about the quality of teamwork from Wikiteams. In *Social computing (SocialCom), 2010 IEEE second international conference* (pp. 17–24). IEEE. doi:10.1109/SocialCom.2010.13.

Viégas, F. B., Wattenberg, M., & Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on human factors in computing systems*. In *CHI '04* (pp. 575–582). New York, NY, USA: ACM. doi:10.1145/985692.985765.

Voss, J. (2006). Collaborative thesaurus tagging the Wikipedia way. *CoRR*. arXiv:cs/0604036

Vuong, B.-Q., Lim, E.-P., Sun, A., Le, M.-T., Lauw, H. W., & Chang, K. (2008). On ranking controversies in Wikipedia: Models and evaluation. In *Proceedings of the 2008 international conference on Web search and data mining*. In *WSDM '08* (pp. 171–182). New York, NY, USA: ACM. doi:10.1145/1341531.1341556.

Wang, D., Zhu, S., & Li, T. (2013). Sumview: A Web-based engine for summarizing product reviews and customer opinions. *Expert Systems with Applications, 40*(1), 27–33. doi:10.1016/j.eswa.2012.05.070.

Wawer, A., & Nielek, R. (2008). Application of automated sentiment extraction from text to modeling of public opinion dynamics. *Polish Journal of Environmental Studies, 17*(3B), 508–513.

Wawer, A., & Nielek, R. (2009). Semantics and the Polish stock market. Towards automated financial Web mining. *Polish Journal of Environmental Studies, 18*, 282–285.

Whiting, S., & Jose, J. M. (2014). Recent and robust query auto-completion. In *Proceedings of the 23rd international conference on World Wide Web*. In *WWW '14* (pp. 971–982). New York, NY, USA: ACM. doi:10.1145/2566486.2568009.

Wierzbicki, A., Turek, P., & Nielek, R. (2010). Learning about team collaboration from Wikipedia edit history. In *Proceedings of the 6th international symposium on Wikis and open collaboration*. In *WikiSym '10* (pp. 27:1–27:2). New York, NY, USA: ACM. doi:10.1145/1832772.1832806.

Yamamoto, Y. (2012). Disputed sentence suggestion towards credibility-oriented Web search. In *Web technologies and applications* (pp. 34–45). Springer.

Yardi, S., & Boyd, D. (2010). Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of Science, Technology & Society, 30*(5), 316–327.

Yasseri, T., Spoerri, A., Graham, M., & Kertész, J. (2014). The most controversial topics in Wikipedia: A multilingual and geographical analysis. In P. Fichman, & N. Hara (Eds.), *Global Wikipedia: International and cross-cultural issues in online collaboration*. Scarecrow Press *arXiv:1305.5566*.

Yasseri, T., Sumi, R., Rung, A., Kornai, A., & Kertész, J. (2012). Dynamics of conflicts in Wikipedia. *PLoS ONE, 7*(6), e38869.

Yu, J., Thom, J. A., & Tam, A. (2007). Ontology evaluation using Wikipedia categories for browsing. In *Proceedings of the sixteenth ACM conference on conference on information and knowledge management*. In *CIKM '07* (pp. 223–232). New York, NY, USA: ACM. doi:10.1145/1321440.1321474.